広島経済大学研究論集
第38巻第3号　2015年12月

大学生のTwitterアカウントの自動検出

石　野　亜　耶*

1. はじめに

近年，Twitter[1)] などのソーシャルメディアが急速な広がりを見せている。Twitterとは，ツイートと呼ばれる140文字のメッセージを投稿できるコミュニティーサイトである。2015年6月30日時点では，Twitterには3億1,600万人月間アクティブユーザがおり，1日平均5億件のツイートが投稿されている[2)]。Twitterには，今日の出来事，食事の内容や交通状況，購入した商品の感想や要望，不満など様々な情報がリアルタイムに投稿されている。現在，これらのTwitter上に投稿された情報を，ビジネスやサービスに役立てるための研究が活発に行われている。例えば，トレンド分析，評判分析，ユーザの属性抽出，実世界の動向（株価や売り上げなど）との相関分析などである［1］。

筆者の担当している授業で調査を行ったところ，約8割の学生がTwitterを利用していることがわかった。詳細は2.3節で述べる。このように大部分の学生が利用していることから，本学でのTwitterの利活用を検討したいと考えている。

大学でのTwitterの活用方法としては，学生が投稿したツイートを解析することで，大学への要望や不満を抽出することが挙げられる。抽出された要望や不満に対して対策を行うことで，学生のよりよい大学生活を支援し，大学に対する満足度を向上させることが可能となる。上記のようなTwitterの利活用を行うためには，本学の学生のTwitterアカウント（以下，学生アカウント）を収集する必要がある。しかし，膨大なTwitterアカウントから，人手で学生アカウントを収集することは困難である。そこで本研究では，学生アカウントを，機械学習を利用し自動で検出する手法を提案する。

本論文の構成は以下の通りである。2章ではTwitterの基本情報と利用状況，3章では関連研究，4章では提案手法，5章では実験と結果について述べる。6章では結論を述べる。

2. Twitterの基本情報と利用状況

2.1節ではTwitterの基本情報について簡単に説明する。2.2節ではTwitterの国内での利用状況，2.3節では本学での利用状況について説明する。

2.1 Twitterの基本情報

筆者は，担当している「Webマイニング」の授業で利用するために，Twitterアカウント（ユーザ名：Aya Ishino）を開設している。図1は，筆者のアカウントのホーム画面である。この図を利用して，Twitterで利用される用語[3)]について簡単に説明する。

①ツイート：Twitterに投稿されるメッセージのこと。最大140文字までのテキストや画像，動画を含めることができる。
②フォロー：Twitterアカウントのツイートが配信されるように登録すること。
③フォロワー：自分のアカウントをフォローしているTwitterアカウントのこと。
④名前：個人的なIDのこと。事業名または

* 広島経済大学経済学部助教

図１　筆者のアカウントのホーム画面

本名の場合もある。

⑤ユーザ名：Twitter でアカウントを識別するためのユーザ特有の ID のこと。

⑥自己紹介：最大160文字で記載できる Twitter アカウントに関する説明のこと。

⑦場所：住んでいる住所や位置のこと。

2.2　国内での **Twitter** の利用状況

本節では，総務省が公開している平成27年度情報通信白書[4] を基に，国内での Twitter の利用状況について説明する。情報通信白書には，日本の情報通信の現状及び情報通信政策の動向などがまとめらており，Twitter を含むソーシャルメディアの利用状況も報告されている。調査方法は，「社会課題解決のための新たな ICT サービス・技術への人々の意識に関する調査研究」（アンケート概要）[5] に記載されている。図２は Twitter などのソーシャルメディアの利用率，図３は Twitter の年代別利用率である。図２と３は，平成 27 年版情報通信白書「第４章第２節ソーシャルメディアの普及がもたらす変化」[6] より作成した。

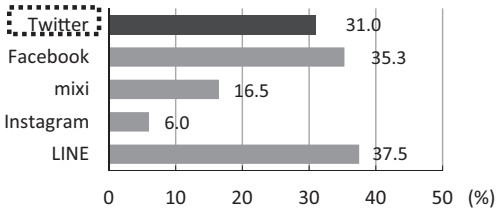

図２　ソーシャルメディアの利用率

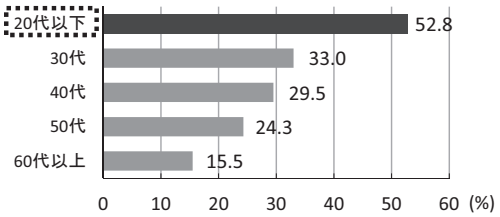

図３　Twitter の年代別利用率

図２より，Twitter の国内での利用率は31.0％である。図３より，20代以下の利用率は52.8％であり，30代以降では徐々に利用率が下がっている。Twitter は，他の年代と比較すると，本学の学生の大部分が含まれる20代以下の利用率が高いソーシャルメディアであることがわかる。

2.3　本学での **Twitter** の利用状況

本節では，本学での Twitter の利用状況を調査した結果を述べる。筆者が担当している「情報社会論」の授業では，受講者のソーシャルメディアの利用状況についてアンケート調査を行っている。この調査は平成26年度の第１回目の授業（2014年４月13日月曜日４限）と平成27年度の第１回目の授業（2015年９月30日水曜日４限）で行った。この調査のうち，Twitter の利用率と利用頻度ついての調査結果を報告する。

Twitter の利用率を調査するため，「Twitter を利用していますか？」という問いに対し，「１：閲覧も投稿もする」，「２：閲覧のみ」，「３：（過去に）利用したことがある」，「４：利用したことがない」のうち１つを選択してもらった。「１：閲覧も投稿もする」または「２：

閲覧のみ」を選択した学生が利用者ということになる。有効回答数は，平成26年度は125件，平成27年度は65件であった。調査結果を図 4 に示す。図 4 からわかるように，利用率は，平成26年度よりも平成27年度の方が増加しており，約 8 割の学生が現在 Twitter を利用している。国内の20代以下の利用率を上回っている。

Twitter の利用頻度を調査するため，「Twitter の利用頻度を教えてください。」という問いに対し，Twitter の利用者（「Twitter を利用していますか？」という問いに対し，「１：閲覧も投稿もする」または「２：閲覧のみ」を選択した学生）に利用頻度を選択してもらった。有効回答数は，平成26年度は86件，平成27年度は46件であった。調査結果を図 5 に示す。実際には，図 5 の区分よりも詳細な利用頻度の調査を行ったが，人数の分布を考慮し，「１：１日に５回以上」，「２：１日に１～４回」，「３：１週間に１～５回」，「４：それ以下」に結果を別けグラフ化した。図 5 からもわかるように，利用頻度

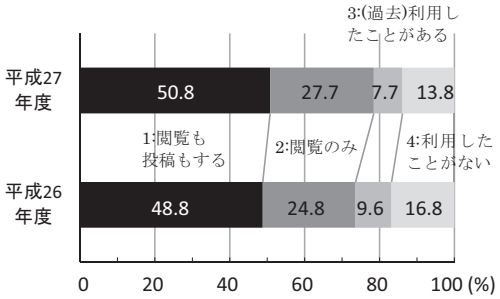

図 4　本学の学生の Twitter の利用率

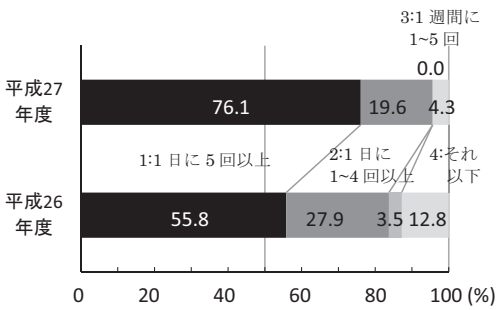

図 5　本学の学生の Twitter の利用頻度

は，平成26年度よりも平成27年度の方が増加しており，９割以上の学生は，１日に１回以上 Twitter を利用している。

以上の調査結果から，本学の学生の Twitter の利用率と利用頻度は高く，平成26年度との比較ではあるが，増加していることがわかる。今後も活発に利用される可能性が高く，大学内での利活用も大いに期待できる。

3. 関 連 研 究

本研究では，学生アカウントを検出する手法を提案している。提案手法は，ユーザの属性推定の一種である。効果的なマーケティングを行う上で重要であることから，ソーシャルメディア上のユーザの属性を推定する研究は活発に行われている。例えば，Twitter アカウントの性別を推定する研究［2］，位置を推定する研究［3, 4］, 任意の種類の属性の推定を行う研究［5］などが挙げられる。本研究では，本学での活用を目的とし，属性を本学に所属している学生かどうかに特化した推定を行う。特定の所属に特化した推定を行うことで，細かな素性の設計ができるため，高い精度で学生アカウントの検出が可能である。

学生アカウントの検出後には，学生アカウントが発信しているツイートから，要望表現を抽出する手法［6］を利用することで，本学に対する要望や不満を抽出する予定である。

大学生の Twitter アカウントに関連する研究について述べる。所属する大学の関係者の Twitter アカウントを人手で収集し，フォロー・フォロワー関係［7］や返信行動［8］に着目し，学部，学科，研究室などの大学内の所属にクラスタリングする手法が提案されている。本研究で検出した学生アカウントに対し，これらの手法を利用することで，大学内での所属を推定することができると考えられる。

大学生の Twitter の利用者が増えるなか，自

身の違法行為やアルバイト等で得た情報をTwitterで発信したことが原因で，批判が集中することで炎上し，無期停学や内定取り消しの処分を下された例もある。ネット炎上分析を行った研究［9］では，大学生はネット炎上のリスクが非常に高いと指摘している。学生アカウントのツイートを分析することで，情報リテラシー教育に役立てることも可能である。

4.　学生アカウントの自動検出手法

学生アカウントの自動検出手法は，以下の３つのステップに分かれている。Step1については4.1節，Step2については4.2節，Step3については4.3節で説明する。

- Step1: 本学に関連するTwitterアカウントの収集
- Step2: 学生アカウントの候補の収集
- Step3: 学生アカウントの自動検出

4.1　本学に関連するTwitterアカウントの収集

膨大なTwitterアカウントのデータを収集し，その中から，学生アカウントを検出するのは困難である。そのため，まずはTwitterのアカウント検索を利用し，本学と関連度の高いTwitterアカウントを収集する。アカウント検索では，名前にキーワードを含むTwitterアカウントを検索することができる。本学に関連するTwitterアカウントを収集するため，使用するキーワードは「広島経済大学」とする。

アカウント検索を利用し収集できたTwitterアカウントの一覧を表１に示す。ツイート数やフォロワー数が少ないTwitterアカウントは，アカウントを開設したものの利用されていない可能性が高い。そのため，表１に示すTwitterアカウントから，フォロワー数が０件である「広島経済大学（非公式）」，ツイート数が０件である「広島経済大学　学生課」，「広島経済大学」，「広島経済大学48期大学祭実行委員会」を

表１　アカウント検索により収集されたTwitterアカウント一覧
（2015年９月21日18：00 最終アクセス）

名前	ユーザ名	ツイート数	フォロー数	フォロワー数
広島経済大学陸上競技部長距離（山猿）	yamazaru_hue	380	117	402
広島経済大学水泳部	HUEST_1	122	59	69
広島経済大学 TRUST	hueflyingdisc	114	92	200
広島経済大学 DANCE RAZZLE	DanceRazzle	47	92	66
広島経済大学 バレー部	HUEvolleyball	35	41	29
広島経済大学陸上競技部	TandF_hue	28	38	97
広島経済大学　大学祭実行委員会	huefes_45th	17	69	127
広島経済大学（非公式）	HUE_hirokei_S42	16	1	0
広島経済大学学友会文化局	bunnkakyoku	14	41	33
広島経済大学 ソフトボール部	huesoftball	13	11	35
広島経済大学 ぶらり安佐南プロジェクト	hue_burari	5	101	49
広島経済大学　学生課	gakuseika	0	2	9
広島経済大学	2010*****	0	0	14
広島経済大学48期大学祭実行委員会	keidaisai48	0	0	4

除いた10件を，本学に関連する Twitter アカウントとして利用する。

4.2　学生アカウントの候補の収集

Step1 で収集した Twitter アカウントは，本学に関連するアカウントであるため，そのフォローやフォロワーは，本学の学生である可能性が高い。そのため，Step1 で収集した Twitter アカウントのフォロー・フォロワーを，学生アカウントの候補として収集する。学生アカウントの候補となるアカウントリストと，アカウント情報は，Twitter Developers[7] から提供されている Twitter API[8] を使用することで収集することができる。収集したデータの一例として，図 1 に示した筆者のアカウント情報を図 6 に示す。このような方法で，10件の本学に関連する Twitter アカウントのフォロー・フォロワーである848件の Twitter アカウントの情報を自動で収集する。

	2998399196
description	広島経済大学の Web マイニングの授業で使用する予定のアカウントです。
statusesCount	207
followersCount	38
favoritesCount	15
friendsCount	22
url	NA
name	Aya Ishino
created	2015/1/27 13:39
protected	FALSE
verified	FALSE
screenName	AyaIshino
location	広島県 広島市 安佐南区
lang	ja
id	2998399196
listedCount	0
followRequestSent	FALSE
profileImageUrl	http://pbs.twimg.com/profile_images/560079374414663680/AfP96_7R_normal.jpeg

図 6　Twitter API を利用して収集したアカウント情報の例

4.3　学生アカウントの自動検出

Step2 で収集した学生アカウントの候補となる Twitter アカウントには，本学の学生のものではないアカウントも含まれている。そのため，Step2 で収集した Twitter アカウントから，学生アカウントを検出する必要がある。本研究では，機械学習を利用することで，大学生の Twitter アカウントを自動で検出する手法を提案する。機械学習にはサポートベクトルマシン（Support Vector Machine, SVM）を使用する。SVM とは，1995年に C. Cortes と V. N. Vapnik によって提案されて以来［10］，その予測精度の高さから爆発的に人気を得た機械学習法であり，自然言語処理においても様々な問題に適用されている学習法の 1 つである。

機械学習に与える素性について説明する。人手で学生アカウントであると判定されたアカウントの自己紹介の例を図 7 に示す。

（例 1）春から広島経済大学(´▽`)
（例 2）HUE 経済学科一年　アーチェリー

図 7　学生アカウントの自己紹介の例

図 7 に示す例のように，学生アカウントの自己紹介には，大学名や学科名が含まれている場合が多い。また，場所が祇園や安佐南区などの大学所在地に設定されている場合もある。よって本研究では，機械学習に以下の素性を使用することで，学生アカウントの自動検出を行う。単語分割は，MeCab[9] により行う。

- ●単語：単語の有無
- ●大学：広島経済大学，広経大，HUE など大学名に関連する単語の有無
- ●学科：ビジネス情報，ビ情など本学の学科に関連する単語の有無
- ●場所：祇園，安佐南など大学所在地に関連する単語の有無

5. 実　　験

４章で述べた提案手法の有効性の確認するため，実験を行った。

5.1 実験手法

データセット

実験用データとして，4.2節で収集したTwitterアカウント848件に対し，アカウント情報を閲覧し，学生アカウントかどうかを人手で判定した結果を用いる。人手で判定を行った結果を表２に示す。

表２　Twitterアカウントの人手での判定結果

学生アカウントである	学生アカウントではない	合計
230	618	848

比較手法

4.2節で収集したTwitterアカウント848件を，全て学生アカウントであると判定した場合をベースラインとした。提案手法として，4.3節で提案した素性を組み合わせた手法１～４で実験を行った。各手法で利用した素性を，表３に示す。

表３　提案手法に用いた素性

提案手法	素性の組み合わせ			
	単語	大学	学科	場所
手法１	○			
手法２	○	○		
手法３	○	○	○	
手法４	○	○		○

機械学習と評価尺度

学生のTwitterアカウントの自動検出の機械学習にはTinySVM[10]を用いた。２次の多項式カーネルを使用し，４分割交差検定を行った。評価尺度として，以下に示す精度・再現率を用いた。精度は検出誤りの少なさを表す評価指標，再現率は検出洩れの少なさを表す評価指標である。現在よりも多くの学生アカウントを検出するためには，提案手法により検出された学生アカウントのフォロー・フォロワーに対し，さらにモデルを適用する方法が考えられる。そのためには，現段階では，検出誤りが少ないほうがよいと考えられる。よって本研究では，再現率よりも精度を重要視する。

$$\text{精度} = \frac{\text{システムが検出した正解件数}}{\text{システムが検出した件数}}$$

$$\text{再現率} = \frac{\text{システムが検出した正解件数}}{\text{人手で判定した正解件数}}$$

5.2 実験結果と考察

実験結果を表４に示す。表４の実験結果より，ベースラインに比べ，提案手法では高い精度を得ることができた。手法３では，素性に学科を加えることで，精度・再現率が低下した。これは，経済学科や経営学科は他大学にもあるためであると考えられる。提案手法では，素性に単語，大学，場所を利用する手法４が最も精度が高かった。本研究では，再現率よりも精度を重要視するため，手法４が最も性能が優れている。今後は，手法４で作成したモデルを使用して，対象とするTwitterアカウントを広げて学生アカウントの検出を行う予定である。

表４　学生のアカウントの自動検出結果

手法（素性）	精度	再現率
ベースライン	0.267	1.000
手法１（単語）	0.897	0.793
手法２（単語＋大学）	0.949	0.954
手法３（手法２＋学科）	0.929	0.934
手法４（手法２＋場所）	0.952	0.941

6. ま と め

本研究では，学生アカウントを自動で検出する手法を提案した。提案手法は，以下の 3 つのステップに分かれている。

- Step1: 本学に関連する Twitter アカウントの収集
- Step2: 学生アカウントの候補の収集
- Step3: 学生アカウントの自動検出

提案手法の有効性を確認するために，実験を行った。実験の結果，精度0.952，再現率0.941という結果を得ることができ，提案手法の有効性を確認することができた。

今後の課題としては，提案手法により検出できた学生アカウントのツイートを収集し，大学に対する要望や不満を抽出することが挙げられる。また，提案手法により検出された学生アカウントのフォロー・フォロワーに対し，さらに作成したモデルを適用することで，多くの学生アカウントを検出する予定である。

注

1） https://twitter.com/
2） https://about.twitter.com/ja/company
3） https://support.twitter.com/categories/281/28
4） http://www.soumu.go.jp/johotsusintokei/whitepaper/index.html
5） http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nd160000.html
6） http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/n4200000.pdf
7） https://dev.twitter.com/
8） https://dev.twitter.com/overview/api
9） http://mecab.sourceforge.net/
10） http://chasen.org/~taku/software/TinySVM/

参 考 文 献

[1] 奥村 学，“マイクロブログマイニングの現在”，電子情報通信学会第 3 回集合知シンポジウム，2012.
[2] John D. Burger, John Henderson, George Kim and Guido Zarrella, “Discriminating gender on Twitter”, Proc. of the EMNLP' 11, pp. 1301–1309, 2011.
[3] Zhiyuan Cheng, James Caverlee and Kyumin Lee, “You are where you tweet: A content-based approach to geo-locating twitter users”, Proc. of the CIKM' 10, pp. 759–768, 2010.
[4] Jeffrey McGee, James Caverlee and Zhiyuan Cheng, “Location prediction in social media based on tie strength”, Proc. of the CIKM' 13, pp. 459–468, 2013.
[5] 上里和也，浅井洋樹，奥野峻弥，山名早人，“Twitter ユーザを対象とした属性推定の精度向上―周辺ユーザの属性補完を利用して―”，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM2015），2015.
[6] Hiroshi Kanayama and Tetsuya Nasukawa, “Textual Demand Analysis: detection of users' wants and needs from opinions”, Proc. of the COLING' 08, Vol. 1, pp. 409–416, 2008.
[7] 畑本典宣，黒澤義明，目良和也，竹澤寿幸，“マイクロブログにおけるユーザのクラスタリングとそのクラスタの特徴語抽出”，言語処理学会第17回年次大会，2011.
[8] 黒澤義明，竹澤寿幸，“マイクロブログサービスの返信行動に着目した投稿及びユーザの分類”，言語処理学会 第17回年次大会，2011.
[9] 田代光輝，“大学生のネット炎上分析と予防及び対応の提案：好ターゲットとしての大学生の実情とネット炎上からの回避の提案”，大妻女子大学紀要，社会情報系，社会情報学研究 21，pp. 233–241，2012.
[10] Corinna Cortes, Vladimir Vapnik, “Support-Vector Networks”, Journal of Machine Learning, pp. 273–297, 1995.